Predicting Information Diffusion in Social Networks using Content and User's Profiles

Cédric Lagnier[1], Ludovic Denoyer[2], Eric Gaussier[1], Patrick Gallinari[2]

[1]Université Grenoble 1, LIG, Grenoble, France
{cedric.lagnier,eric.gaussier}@imag.fr
[2]Université Pierre et Marie Curie, LIP6, Paris, France
{ludovic.denoyer,patrick.gallinari}@lip6.fr

Abstract. Predicting the diffusion of information on social networks is a key problem for applications like Opinion Leader Detection, Buzz Detection or Viral Marketing. Many recent diffusion models are direct extensions of the *Cascade* and *Threshold* models, initially proposed for epidemiology and social studies. In such models, the diffusion process is based on the dynamics of interactions between neighbor nodes in the network (the social pressure), and largely ignores important dimensions as the content of the piece of information diffused. We propose here a new family of probabilistic models that aims at predicting how a content diffuses in a network by making use of additional dimensions: the content of the piece of information diffused, user's profile and willingness to diffuse. These models are illustrated and compared with other approaches on two blog datasets. The experimental results obtained on these datasets show that taking into account the content of the piece of information diffused is important to accurately model the diffusion process.

1 Introduction

The emergence of Social Networks and Social Media sites has motivated a large amount of recent research. Different problems are currently studied such as social network analysis, social network annotation, community detection, link prediction or information diffusion. Many recent information diffusion models are extensions of the widely used independent cascade model (IC) [5] and linear threshold model (LT) [6], and view diffusion as an iterative process in which the probability of diffusion depends, for each user, on her incoming neighbors having already diffused the information. However, while IC or LT inspired models can be used for this task they suffer from two main drawbacks:

– They do not consider the content of the piece of information to be diffused, while this seems an important factor: for the same network, two different pieces of information will propagate differently depending on the respective fields of interest of the different users in the network;

– They do not consider any information about the users of the social networks, as user profiles for example, while this information is intuitively very informative for characterizing how much and how a user tends to diffuse a message.

In this study, we introduce a new family of diffusion models that (a) make use of the content of the information diffused, (b) take into account the profile of each user as well as (c) their willingness to diffuse a given piece of information. Experiments for assessing the validity of this new family of models are performed on two real, widely used datasets extracted from the blogosphere.

The remainder of the paper is organized as follows. Section 2 introduces the notations used throughout this study and states the problem addressed. Section 3 describes the different features used, while Section 4 presents the probabilistic models built on top of these features. These models are evaluated and compared to standard information diffusion models in Section 5. Lastly, Section 6 describes the related work, while Section 7 concludes the study.

2 Notations and Problem Statement

We consider here a social network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ composed of a set of nodes or users $\mathcal{N} = \{n_1, ..., n_N\}$ and a set of directed edges $\mathcal{E}$. We denote by $\mathcal{B}(n_i)$ the set of nodes with an incoming link to n_i such as $\forall n_j \in \mathcal{B}(n_i), (n_j, n_i) \in \mathcal{E}$. Elements of $\mathcal{B}(n_i)$ will be called "incoming neighbors" of n_i ($|\mathcal{B}(n_i)|$ denotes the cardinal of $\mathcal{B}(n_i)$) and the set of incoming neighbors of n_i having already diffused content c^k before or at time t will be denoted $\mathcal{Q}^k(n_i, t)$. We furthermore assume that we have access to:

– $\mathcal{C} = (c^1, ..., c^K)$, the set of contents diffused through the network. c^k is a vector of features representing the content diffused.
– $\mathcal{P} = (p^1, ..., p^N)$, the set of user profiles; p^i is a vector of features representing the interests of user n_i and is defined on the same feature space as the one used for $\mathcal{C}$; Such vectors can directly be inferred from the contents diffused in the past by users, as the posts in blogs for example;
– $\mathcal{M} = (M^1, ..., M^K)$, a set of diffusion matrices where $m^k_{i,t} \in \{0, 1\}$; $m^k_{i,t} = 1$ indicates that user n_i has diffused content c^k before or at time t. Such a user will also be called a **contaminated user** in the following. T corresponds to the duration of all diffusions, in time steps. $M^k_{.,t}$ will denote the t^{th} column of M^k. Lastly, the set $\mathcal{M}$ is divided into two disjoint subsets: a set of training matrices, $\mathcal{D} = \left\{(M^1, c^1), ..., (M^\ell, c^\ell)\right\}$, and a set of test matrices, $\mathcal{T} = \left\{(M^{\ell+1}, c^{\ell+1}), ..., (M^K, c^K)\right\}$. Training matrices will be used to learn diffusion models, whereas test matrices will be used for evaluation.

We are interested here in the step-by-step evolution of the diffusion process, as well as in its result after a given time. We denote by $\mathcal{F}_s$ the function that predicts the diffusion of an information at time t given the diffusion status of the network at time $t - 1$. With the elements defined above:

$$m^k_{i,t} = \mathcal{F}_s(n_i, \mathcal{G}, \mathcal{P}, c^k, M^k_{.,t-1}) \tag{1}$$

The function $\mathcal{F}_g$ predicting the result of the diffusion process after a given time can be constructed from $\mathcal{F}_s$ by "unfolding" it over time: $\mathcal{F}_g(n_i, t, \mathcal{G}, \mathcal{P}, c^k, M^k_{.,0}) = \mathcal{F}_s^{(t)}(n_i, \mathcal{G}, \mathcal{P}, c^k, M^k_{.,0})$, where $^{(t)}$ denotes the composition of $\mathcal{F}_s$ t times. In previous studies, $\mathcal{F}_s$ depends neither on $\mathcal{P}$ nor on c^k, and we make here the assumption that exploiting information from $\mathcal{P}$ and c^k will result on a better prediction of how information diffuses.

The goal of the present study is thus twofold:

1. Learn, from $\mathcal{G}$, $\mathcal{P}$, c^k and the training set $\big((M^1, c^1), ..., (M^\ell, c^\ell)\big)$, the mapping $\mathcal{F}_s$;
2. Assess whether exploiting $\mathcal{P}$ and c^k leads to better diffusion models.

3 A User-based Approach

We show in this section how the different aspects mentioned can be captured through simple feature functions.

The **thematic interest** of each user in the content diffused can be modeled as a proximity between user profiles (describing their interests) and the content diffused. A general form for this proximity is:

$$S(n_i, \mathcal{P}, c^k, \theta_s) = sim(p^i, c^k) - \theta_s$$

where θ_s is a threshold and $sim(p^i, c^k)$ represents a similarity between the content diffused and the user profile. Setting θ_s to 0 amounts to relying solely on the similarity between the user profile and the content diffused; higher values of θ_s allow one to "discourage" diffusion when the user interest in the content is not sufficient. We use in this study the *cosine* similarity for *sim*, but other choices are possible.

The **activity**, or active/passive role, can directly be measured, on the training set, through the ratio between the number of contents received and diffused by a user and the number of contents received by that user:

$$Act(n_i, \mathcal{G}, \mathcal{D}) = \frac{\sum_{k=1}^{l} I(|\mathcal{Q}^k(n_i, T-1)| > 0) m^k_{i,T}}{\sum_{k=1}^{l} I(|\mathcal{Q}^k(n_i, T-1)| > 0)}$$

where $I()$ denotes the indicator function. This measure can be generalized by introducing a threshold, through:

$$W(n_i, \mathcal{G}, \mathcal{D}, \theta_w) = Act(n_i, \mathcal{G}, \mathcal{D}) - \theta_w$$

$W(n_i, \mathcal{G}, \mathcal{D}, \theta_w)$ represents the willingness of user n_i to diffuse information, and θ_w plays a role similar to the one of θ_s above.

Lastly, the **social pressure** on each user, i.e. the fact that many different neighbors have diffused a given content, is traditionally measured, either implicitly or explicitly, through the number of incoming neighbors having already diffused the information. We denote the associated measure:

$$SP(n_i, \mathcal{G}, M^k, t)$$

The particular form this measure takes depends on the model retained, and will be detailed in Section 4.

Each user can thus be represented by a vector of three features evolving over time for each content c^k, a vector we denote Φ^{n_i,t,c^k}, omitting, for readability reasons, the other arguments $(\mathcal{P}, c^k, \mathcal{G}, M^k_{.,T-1}, \theta_s, \theta_w)$:

$$\Phi^{n_i,t,c^k} = \begin{pmatrix} S(n_i, \mathcal{P}, c^k, \theta_s) \\ W(n_i, \mathcal{G}, \mathcal{D}, \theta_w) \\ SP(n_i, \mathcal{G}, M^k, t) \end{pmatrix}$$

These features are then combined through simple linear combinations to yield basis functions for each user, content and time step:

$$f_\lambda(n_i, t, c^k) = \lambda_0 + \lambda_1 \Phi_1^{n_i,t,c^k} + \lambda_2 \Phi_2^{n_i,t,c^k} + \lambda_3 \Phi_3^{n_i,t,c^k} \tag{2}$$

where $\lambda_0, \cdots, \lambda_3$ are parameters that need to be learned. The way $\mathcal{F}_s$ and $\mathcal{F}_g$ are constructed from the basis functions f_λ will be detailed in section 4.

4 Probabilistic modeling

Probabilistic models for information diffusion allows one to model the uncertainty inherent to the diffusion process. In this case, one does not consider that each user has either diffused a given content or not, but rather that each user has a certain probability of having diffused the given content. Two quantities are useful here: $P(n_i, c^k, t)$, the probability that user n_i diffuses content c^k at time t, and $P(n_i, c^k, \leq t)$, the probability that user n_i has diffused content c^k before time t. These two quantities are related through:

$$P(n_i, c^k, \leq t+1) = P(n_i, c^k, \leq t) + (1 - P(n_i, c^k, \leq t))P(n_i, c^k, t) \tag{3}$$

A user having diffused before time $t+1$ has either diffused before time t, or has not and has diffused at time t. Furthermore, because of the definition of $P(n_i, c^k, \leq t)$:

$$\mathcal{F}_s(n_i, t, \mathcal{G}, \mathcal{P}, c^k, M^k_{.,t-1}) = P(n_i, c^k, \leq t)$$

and $\mathcal{F}_g$ can be obtained by unfolding the process over time, i.e. computing $\mathcal{F}_s$ from $t = 0$ to the desired time.

When the thematic interest of the user is high, or when her willingness to diffuse or her social pressure is high, $P(n_i, c^k, t)$ should be high; conversly, when thematic interest, willingness to diffuse and social pressure are low, $P(n_i, c^k, t)$ should be low. Such a behavior is naturally captured in the logistic function, which acts as a soft thresholding process and yields valid probability functions. Furthermore, a user cannot diffuse a content if no incoming neighbor has already diffused it. Because of the probabilistic setting retained here, one does not have a direct access to $|\mathcal{Q}^k(n_i, t)|$, the number of incoming neighbors having already diffused, but rather to an expectation of it ($E[|\mathcal{Q}^k(n_i, t)|]$). Hence:

$$SP(n_i, \mathcal{G}, M^k, t) = E[|\mathcal{Q}^k(n_i, t)|]$$

and:

$$P(n_i, c^k, t) = \begin{cases} (1 + e^{-f_\lambda(n_i,t,c^k)})^{-1} \text{ if } E[|\mathcal{Q}^k(n_i,t)|] > 0 \\ \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

with (λ_1, λ_2, λ_3) positive or null (when a feature has no impact on the diffusion).

The expectation $E[|\mathcal{Q}^k(n_i,t)|]$ is defined as $\sum_{m=0}^{|\mathcal{B}(n_i)|} m\ P(|\mathcal{Q}^k(n_i,t)| = m)$, where $P(|\mathcal{Q}^k(n_i,t)| = m)$ is the probability that the number of incoming neighbors who have diffused the content is m. It is easy to show that (we skip here the derivation which is purely technical):

$$E[|\mathcal{Q}^k(n_i,t)|] = \sum_{n_j \in \mathcal{B}(n_i)} P(n_j, c_k, \leq t) \quad (5)$$

The dynamics of the diffusion thus evolves, from one time step to another, through:

1. Initialization: $P(n_i, c^k, \leq 0) = 1$ for initial diffusers, 0 otherwise;
2. Iteratively compute (from $t = 0$):
   - $E[|\mathcal{Q}^k(n_i,t)|]$ using equation 5
   - $P(n_i, c^k, t)$ using equation 4
   - $P(n_i, c^k, \leq t+1)$ using equation 3

The main problem with the above model, however, is that the probabilities $P(n_i, c^k, \leq\ t)$ cannot decrease, and will necessarily increase if $P(n_i, c^k, t)$ is strictly positive at some point in time. This is due to the fact that users are "aware" of the content they have already diffused at all time steps, and that their probability of diffusing will be reinforced by subsequent receptions of a given content (for this reason, we refer to this model as **RUC**, for *Reinforced User-Centric*). The following model corrects this drawback.

A time-decaying extension The quantity $P(n_i, c^k, t)$ becomes strictly positive as soon as $E[|\mathcal{Q}^k(n_i,t)|]$ is strictly positive, and one would like, in this latter measure, that the influence of users having diffused an information a long time ago be less important than the one of users having diffused the information recently. One can thus replace equation 5 by the following equation:

$$E[|\mathcal{Q}^k(n_i,t)|] = \sum_{n_j \in \mathcal{B}(n_i)} \rho(n_j, c_k, t) \quad (6)$$

where $\rho(n_j, c_k, t)$ is a function of the influence n_j has on her outgoing neighbors at time t wrt content c_k, penalizing "old" diffusions:

$$\rho(n_j, c_k, t+1) = \delta \times \rho(n_j, c_k, t) + (1 - P(n_j, c_k, \leq t))P(n_j, c_k, t) \quad (7)$$

By definition, $\rho(n_j, c_k, t = 0) = 1$ for initial diffusers and 0 otherwise. $\delta,\ 0 \leq \delta \leq 1$ is a decaying parameter controlling the penalization on old diffusions. When $\delta = 1$, $\rho(n_j, c_k, t) = P(n_j, c_k, \leq t)$ and one recovers the RUC model. The other quantities of the RUC model remain unchanged. We will refer to the model with a decaying parameter as **DRUC**, for Decaying Reinforced User-Centric.

Setting θ_s and θ_w We now turn to the problem of setting the thresholds θ_s and θ_w. A user having a similarity with the content above θ_s is more likely to diffuse an information; conversely, a user with a similarity below θ_s is more likely to not diffuse the information. The global similarity function defined above is positive in the first case and negative in the second one. θ_s thus corresponds to a threshold on the similarity function above which a user is more likely to diffuse an information, and can be obtained, from the training set, through a line search on the cosine values between content diffused and user profiles. This line search process is here initialized at 0, with an increment of 0.05, and is stopped as soon as the number of users re-diffusing a content is greater than the number of users not re-diffusing it. A similar reasoning for the willingness to diffuse ($W(n_i, \mathcal{G}, \mathcal{D}, \theta_w)$) directly leads to $\theta_w = 0.5$.

Estimating the λs The parameters ($\lambda_0, \lambda_1, \lambda_2$ and λ_3) can be learned through maximum likelihood, with positivity constraints. Let $\mathcal{L}(\lambda_0, \lambda_1, \lambda_2, \lambda_3)$ denote the likelihood of the training set. The learning problem can be formulated as:

$$\begin{cases} \text{argmax}_{\lambda_0,\lambda_1,\lambda_2,\lambda_3} \mathcal{L}(\lambda_0, \lambda_1, \lambda_2, \lambda_3) \\ \text{subject to: } \lambda_1 \geq 0,\ \lambda_2 \geq 0,\ \lambda_3 \geq 0 \end{cases}$$

and one can resort, to solve this problem, to a projected gradient approach, in which each gradient ascent step is followed by a projection of the parameters on the admissible intervals.

The likelihood, on the training set, for the above models is given by:

$$\mathcal{L}(\lambda_0, \lambda_1, \lambda_2, \lambda_3) = \prod_{k=1}^{l} \prod_{t=1}^{T} [\prod_{n_i \in \mathcal{Q}^k(t)} P(n_i, c^k, \leq t) \prod_{n_i \notin \mathcal{Q}^k(t)} (1 - P(n_i, c^k, \leq t)]$$

where $\mathcal{Q}^k(t)$ is the set of all users having diffused content c^k before time t. For efficiency reasons, we make use of the recurrence equation (Eq. 3) to compute the partial derivatives, and store, for each user, the current values of $P(n_i, c^k, \leq t)$ and its derivatives.

5 Experiments

We compare here the models presented above with several baseline diffusion models used in previous studies. This comparison will help us assessing how much the new dimensions considered in the user-centric family of models are useful for content diffusion. The models we have retained are the following:

1. The **Independent Cascade Model** (IC). Its parameters are learned through the EM algorithm proposed in [19];
2. The **Asynchronous Independent Cascade Model** (ASIC) which is described in [18] and represents an asynchronous version of the IC model.
3. The recently introduced **NetRate** model [17], with the exponential distribution;

Dataset	# nodes	# links	# terms	# cascades	Mean size	Max size
MemeTracker (Dense)	5000	4373	24482	2977	1.21	4
ICWSM (Dense)	5000	17746	173014	23738	1.075	11
MemeTracker (Sparse)	39427	10816	70602	104973	0.006	10
ICWSM (Sparse)	40268	62657	262290	104980	0.018	33

Table 1. Main statistics of datasets for the *Sparse* and *Dense* versions

4. The **RUC** and **DRUC** models presented in Section 4; In this study, we have arbitrarily set the parameter δ to 0.9, which amounts to consider a small decay over time.

In order to compare the different methods, we make use of two datasets:

- The **ICWSM** [3] dataset is composed of blog posts and links between them. Each user corresponds to a blog and diffusion of information is observed through links between blogs: if post p_2 of blog b_2 contains an hyperlink to post p_1 of blog b_1, then we consider that b_2 has diffused the content coming from b_1;
- The **MemeTracker** [10] dataset is composed of blog posts and links between them. Contrary to the ICWSM dataset, no blog url is attached to a post. We thus inferred blogs using post urls (a post url contains the url of the blog it belongs to). To do so, we cut post urls at the first "/" character after "http://" and assume that the string obtained corresponds to the url of the blog. As for the ICWSM dataset, we consider that information propagates from one user (blog) to another if there is a link from a post of the former to a post of the latter.

The graph between blogs is built from the above datasets: two blogs u_i and u_j are connected if at least one information diffuses between u_i and u_j.

For each dataset, we have extracted two different corpora:

- The **Sparse** corpora have been built by selecting randomly 100,000 cascades of blog posts. In this case, many of the selected cascades do not diffuse over the network resulting in a case where the models can only be trained on a few number of diffusions. These corpora are used to evaluate the models in a context of low diffusion.
- The **Dense** corpora have been built by focusing on a subset of the $5,000$ users that are the most active. We have only kept the cascasdes over these active users which have been linked at least one time. These two corpora are used to evaluate the models in the context of a dense diffusion.

The number of users, cascades and the mean size of the cascades are given in Table 1. The length of a cascade is 1 if the information diffuses once from a initial user to another one. As one can see, *Sparse datasets* are composed of low length cascades – i.e. many cascades do not diffuse – while *Dense datasets* are composed of larger cascades. The parameter θ_s has been computed as explained in section 4 and set to 0.35 for MemeTracker datasets and 0.4 for ICWSM datasets.

For each corpus we performed the following normalization operations:

- Taking posts during only one month;
- Filtering out of non-English posts;
- Removal of empty words with empty words list.
- Stemming using Porter stemming;
- Filtering out of words appearing less than five times.

The above preprocessing then yields a standard word vector for each post. The vector for a cascade is then computed by averaging the vectors of all the posts that compose a cascade. The profile of each user is computed by averaging the vectors of the cascades diffused by the user on the training set. In order to evaluate the different models, we use a 5-fold cross validation scheme (4 blocks for training, one for testing). Training blocks are used to estimate models parameters and the last one is used for the evaluation. All the results presented below are averaged over the 5 different splits.

In order to evaluate the quality of the proposed approaches and baseline models, we use a specific precision measure: we compute the *Precision at different Recall Points (PRP)*. This measure computes the precision curves following procedure, for each cascade:

1. The nodes scores (probabilities to be contaminated) obtained with a given model are ordered in decreasing order of their values.
2. Precision is computed at each point of recall - at each rank ℓ where the real contamination score of the user is 1.

PRP values are averaged over all the testing cascades. The precision at the first recall point reflects the ability of a model to find one user that will be contaminated, the second point corresponds to the ability of the model to find two contaminated users,... Note that only the cascades of at least length ℓ are used to evaluated the precision at rank ℓ – i.e. performances on high ℓ values are less robust than estimation made for low ℓ values. This measure has been used previously [14].

Results on Sparse Corpora The PRP values over the two sparse datasets are illustrated in Tables 2 and 3. First, one can see that baseline models (IC, ASIC and Netrate) perform poorly on these datasets. As explained before, this is mainly due to the fact that, on *Sparse* datasets, the number of diffusing training cascades is very low resulting in baseline methods that predict almost no diffusion.

The assumptions made by our approaches are different. Particularly the diffusion of information is modeled through a set of features that is shared by all users. This allows us to transfer the behavior of one user to another instead of learning the behavior of each user separately. Our approach makes the learning problem easier and offers better generalization abilities explaining the higher prediction performances.

	Sparse					Dense			
Cascade length	1	2	3	4	≥5	1	2	3	4
Nb cascades	149	28	9	5	≤4	596	40	7	2
IC	0.02	0.04	0	0	0	0.29	0.20	0.38	0.33
ASIC	0.03	0.07	0	0	0	0.14	0.15	0.32	0.33
Netrate	0.02	0	0	0	0	0.16	0.15	0.27	0
RUC	0.58	0.47	0.36	0.31	≤0.28	**0.63**	**0.50**	**0.63**	**0.67**
DRUC	**0.64**	**0.52**	**0.37**	**0.32**	**≤0.28**	**0.63**	**0.50**	**0.62**	**0.68**

Table 2. Precision values on the **MemeTracker** datasets. The number of cascades used for computing precision at each recall point is illustrated in the second line. Bold indicates best results.

	Sparse					Dense				
Cascade length	1	2	3	4	≥5	1	2	3	4	≥5
Nb cascades	440	88	33	16	≤10	4748	656	255	90	≤18
IC	0.13	0.03	0	0	0	0.73	0.66	0.71	0.72	≤0.21
ASIC	0.07	0	0	0	0	0.36	0.30	0.32	0.35	≤0.03
Netrate	0.03	0.01	0.03	0	0	0.12	0.01	0	0	0
RUC	0.70	0.62	0.61	0.67	≤0.56	0.83	0.75	0.75	0.79	≤0.52
DRUC	**0.73**	**0.66**	**0.64**	**0.72**	**≤0.68**	**0.85**	**0.77**	**0.78**	**0.81**	**≤0.56**

Table 3. Precision values on the **ICWSM** datasets. The number of cascades used for computing precision at each recall point is illustrated in the second line. Bold indicates best results.

Results on Dense Corpora Concerning the *Dense* datasets – Tables 2 and 3 – one can see that baseline models perform better than previously due to the higher number of training cascades that diffuse. The best baseline model is the IC model that clearly outperforms ASIC and Netrate. We think that this is due to the fact that ASIC and Netrate introduce a strong decay in the diffusion through an exponential model. As the number of diffusions in each dataset is still low, the probability predicted by these models is also low and dominated by the decay exponential term (of the form $e^{-P_{ij}(t-t_0)}$). The difference between these values is thus small and the models fail to differentiate between diffusions and non-diffusions.

The improvement provided by RUC and DRUC approaches is particularly important on the *Dense MemeTracker* dataset – at the first point of recall, RUC has a precision of 0.63 where IC only obtains 0.29 – and significant on the *Dense ICWSM* dataset. These results show the importance of considering the three different features, namely *thematic interest*, *activity* and *social pressure*. Furthermore, the values obtained by the parameter of the thematic interest feature (λ_1) are systematically higher, for both RUC and DRUC, and for both *ICWSM* and *MemeTracker* than that obtained for the other parameters (for example, on *MemeTracker*, the values obtained are $\lambda_1 = 7.01$, $\lambda_2 = 5.92$, $\lambda_3 = 2.78$ for RUC, and $\lambda_1 = 9.49$, $\lambda_2 = 3.99$, $\lambda_1 = 0.95$ for DRUC). Even though it is difficult to

compare features on the sole basis of the values taken by their associated parameter, the above values clearly show that the thematic interest plays a crucial role in the information diffusion process (the social pressure becoming a minor player for the DRUC model). This fully justifies our will to take into account the content of the information in the diffusion process. Indeed, the process will be different for different pieces of information, even if the same initial diffusers are used.

Comparison between the different UC models The experiments show that in average, DRUC outperforms RUC on three over four datasets. This is particularly true over the large cascades because the DRUC model is better for modelling long diffusions - see Section 4. Due to the high variance of the results on sparse datasets, the difference between RUC and DRUC is not significant (Wilcoxon test with a p-value of 0.05); it is however significant for the dense ICWSM dataset.

6 Related work

Information diffusion models can roughly be classified into two main categories: contagion models, in which the diffusion is based on a probability of diffusion between users in contact (see for example [5, 15, 8, 9]), and influence models, also called threshold models, in which a user diffuses an information if the number or the proportion of her incoming neighbors who have already diffused the information is above a user-specific threshold (see for example [6, 13, 2]),

The prototype for contagion models is the IC (*Independent Cascade* model, which has recently been extended to integrate a time variable in the diffusion model and to account for the fact that diffusion/contamination can be delayed. To do so, the *ASIC* (Asynchronous IC), introduced in [18], makes use of an exponential probability distribution to model the delay between the contamination of a user and its attempt to contaminate her neighbors, the contamination probability decreasing with this delay (a similar "latence" phenomenon is used in [11]). More recently, [17] consider different probability distributions for the delay in the contamination: exponential, power law and Rayleigh distributions. The family of models thus defined is called *NetRate*. The version based on the exponential distribution is in fact a special case of the ASIC model (obtained when setting the $k_{v,w}$ parameter of ASIC to a constant). In [20], the ASIC model is further enriched with node attributes information, leading to a model that is similar to the probabiistic model presented here. However, this extension allows one to capture the similarity between users through the attributes they share, and does not account for the features we have retained here. In particular, the final model obtained will predict the same diffusion from the same set of initial diffusers, no matter which information is diffused.

The prototype for influence models is the LT (*linear threshold*) model, originally defined in [6], and extended in [7, 13, 21, 1, 16, 4, 12, 2]. In a similar vein,

[22] introduces a linear influence model based on time series and aiming at determining the "volume" of users who have diffused an information after a given time, a task which differs from the one addressed here (as not only the diffusion volume but also the particular users having diffused are searched for). For all these models, however, and similarly to the contagion models, only the social pressure is used to determine the fact that a given user will diffuse an information or not, which radically differs from the setting adopted in this study.

7 Conclusion

We have proposed here a new family of models (*User-Centric models*) that aims at predicting how a content diffuses in a network by making use of three dimensions, namely the content diffused, the users profiles and their willingness to diffuse. In particular, we have shown how to integrate these dimensions into simple feature functions, and proposed a new probabilistic model to take them into account. We have furthermore illustrated and compared our models with other approaches on two blog datasets. The experimental results obtained on these datasets show that (a) the content of the information diffused plays a major role in the diffusion process and should not be ignored, as was done so far, (b) user's profiles also play an importnat role, which was recognized in recent studies on information diffusion even though not systematically used, and (c) state-of-the-art results can be obtained to models relying on few, adequate parameters, as is the case for the models introduced here which make use of only 3 parameters compared to the thousands of parameters used in the IC-based models.

A direct extension of our work would be to deal with various types of content (images, videos, text) and cascading behaviors (small versus long cascades) and predict the diffusion of heterogeneous information. In this study, we have arbitrarily set the decay parameter of DRUC and we project to estimate it in future works. Another extension we plan on addressing is to simultaneously take into account different social networks, so as to escape away from the close-world assumption underlying most of the studies in information diffusion.

Acknowledgements

This work has been partially supported by the ARESOS project from CNRS Program MASTODONS and the DIFAC FUI project

References

1. Abrahamson, E., Rosenkopf, L.: Social network effects on the extent of innovation diffusion: A computer simulation. Organization Science 8, 289–309 (1997)
2. Borodin, A., Filmus, Y., Oren, J.: Threshold models for competitive influence in social networks. October pp. 1–15 (2010)
3. Burton, K., Java, A., Soboroff, I.: The ICWSM 2009 Spinn3r Dataset. In: The Third Annual Conference on Weblogs and Social Media (ICWSM 2009) (May 2009)

4. Dodds, P., Watts, D.: Universal Behavior in a Generalized Model of Contagion. Physical Review Letters 92, 218701 (2004)
5. Goldenberg, J., Libai, B., Muller, E.: Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. Marketing Letters pp. 211–223 (2001)
6. Granovetter, M.: Threshold Models of Collective Behavior. American Journal of Sociology 83, 1420–1443 (1978)
7. Granovetter, M., Soong, R.: Threshold models of diversity: Chinese restaurants, residential segregation, and the spiral of silence. Sociological Methodology 18, 69–104 (1988)
8. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 137–146. ACM Press (2003)
9. Kimura, M., Saito, K., Nakano, R.: Extracting influential nodes for information diffusion on a social network. Proceedings Of The National Conference On Artificial Intelligence 22(2), 1371 (2007)
10. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 497–506. KDD '09, ACM (2009)
11. Liben-Nowell, D., Kleinberg, J.: Tracing information flow on a global scale using Internet chain-letter data. Proceedings of the National Academy of Sciences 105, 4633–4638 (2008)
12. López-Pintado, D., Watts, D.J.: Social Influence, Binary Decisions and Collective Dynamics. Rationality and Society 20, 399–443 (2008)
13. Macy, M.W.: Chains of Cooperation: Threshold Effects in Collective Action. American Sociological Review 56, 730–747 (1991)
14. Manning, C.D., Raghavan, P., Schtze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
15. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
16. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 61–70. KDD '02, ACM (2002)
17. Rodriguez, M.G., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. In: Getoor, L., Scheffer, T. (eds.) Proceedings of the 28th International Conference on Machine Learning. pp. 561–568. ICML '11, ACM (2011)
18. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. Learning 5828, 322–337 (2009)
19. Saito, K., Nakano, R., Kimura, M.: Prediction of information diffusion probabilities for independent cascade model. In: Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III. pp. 67–75. KES '08, Springer-Verlag (2008)
20. Saito, K., Ohara, K., Yamagishi, Y., Kimura, M., Motoda, H.: Learning diffusion probability based on node attributes in social networks. In: Kryszkiewicz, M., Rybinski, H., Skowron, A., Ras, Z.W. (eds.) ISMIS. Lecture Notes in Computer Science, vol. 6804, pp. 153–162. Springer (2011)
21. Valente, T.W.: Network Models of the Diffusion of Innovations (Quantitative Methods in Communication Subseries). Hampton Press (NJ) (1995)
22. Yang, J., Leskovec, J.: Modeling information diffusion in implicit networks. In: IEEE International Conference on Data Mining. Stanford InfoLab (2010)